

## Stat 8053: Ch. 10 Factor Analysis, rev November 20, 2011

A common motivation for factor analysis comes from psychology. Suppose that  $X$  is a  $p \times 1$  vector of observed test scores, sometimes called *manifest* variables, meaning that they can be observed. These could be, for example, final exam scores in several different subjects taken by third grade students. The factor analysis model posits that the manifest variables are actually the result of a few *latent* variables,  $F$ , of dimension  $k \times 1$ . In the simplest case,  $F$  is just a scalar, and corresponds to the construct “intelligence”. Alternatively, intelligence could have more than one dimension, such as “quantitative intelligence” and “verbal intelligence”, so now  $k = 2$ . The value of  $k$  is expected to be small relative to  $p$ , but is otherwise unrestricted.

The basic factor analysis model is given by:

$$x|f = Qf + u + \mu \quad (1)$$

where we condition on  $f$ , and

$x$  is a  $p \times 1$  vector of observed *manifest* variables.

$f$  is a  $k \times 1$  vector of *unobserved*, or *latent* common factor variables. The latent factors are random variables; for identifiability we assume they have mean 0 and covariance matrix  $I$ .

$Q$  is a  $p \times k$  unobserved matrix of *factor loadings* that are like the predictor matrix in a linear regression.

$u$  is an unobservable  $p \times 1$  random vector of *specific factors* assumed to have mean 0 and covariance matrix  $\text{Var}(u) = \Psi = \text{diag}(\psi_1, \dots, \psi_p)$ .

$\mu$  is the unconditional mean  $\mu = E(x) = E_f E(x|f)$ .

According to this model, the manifest variables consist of two parts, one due to a linear combination of the factors in  $f$ , and a second unexplained part in  $u$ . The only observable quantity (1) is  $x$ ; everything on the right is unobservable.

We will further assume that the distributions of  $x$ ,  $f$  and  $u$  are all normal (other versions of factor analysis use other distributions). It then follows that

$$x|f \sim N(Qf + \mu, \Psi)$$

and the unconditional distribution, averaging over the distribution of  $f$  is

$$x \sim N(\mu, \Sigma = Q\text{Var}(f)Q' + \text{Var}(u) = QQ' + \Psi) \quad (2)$$

**Principal Components** If  $\Psi = \sigma^2 I$ , the (1) is the probabilistic principal components analysis we discussed previously, and so we get the PC solution we discussed previously. By allowing  $\Psi$  to be a diagonal matrix we get a larger class of possible solutions.

**Estimation** The only estimates we consider are maximum likelihood, assuming (2). The likelihood was derived in class, and is given in the textbook. The data will consist of the  $n \times p$  matrix of manifest variables  $X$ , each of whose rows satisfies (2). The sufficient statistic, assuming  $X$  is centered and scaled, is the sample correlation matrix, which has  $p(p+1)/2$  unique elements. All parameters of interest are in  $\Sigma$ . The factor loading matrix  $Q$  has  $pk$  parameters for a  $k$ -factor solution, while  $\Psi$  has  $p$  parameters. Additional constraints on the parameters are introduced to get a unique solution, and these introduce an additional  $k(k-1)/2$  parameters (see the textbook for details). Estimation is possible as long as the number of unique elements in the correlation matrix exceeds the number of parameters and constraints.

## US Company Data

We continue with the US Companies data after removing the two unusual companies found in the last handout. We create a new variable called `sector` which represents the type of company, as described in the textbook.

```
> loc <- "http://www.stat.umn.edu/~sandy/courses/8053/Data/uscomp1.dat"
> head(uscomp <- read.table(url(loc),header=TRUE))
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
1	19788	9084	10636	1092.9	2576.8	79.4
2	5074	2557	1892	239.9	578.3	21.9
3	13621	4848	4572	485.0	898.9	23.4
4	1117	1038	478	59.7	91.7	3.8
5	1633	701	679	74.3	135.9	2.8
6	5651	1254	2002	310.7	407.9	6.2

```
> snames <-c("Com", "Enr", "Fin", "HiTch", "Manu", "Med", "Oth", "Ret", "Tran")
> sector <- rep(1:9, c(2,15, 17, 8, 10, 4, 7, 10, 6))
> print(R <- cor(uscomp[-c(38,40),]), digits=3)
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
Assets	1.0000	0.5072	0.415	-0.0388	0.159	0.209
Sales	0.5072	1.0000	0.599	0.0721	0.369	0.796
MarketValue	0.4146	0.5989	1.000	0.4700	0.662	0.648
Profits	-0.0388	0.0721	0.470	1.0000	0.884	0.179
CashFlow	0.1592	0.3695	0.662	0.8841	1.000	0.336
Employees	0.2086	0.7961	0.648	0.1786	0.336	1.000

The decomposition (2) is to be estimated to match the sample correlation matrix  $R$  as closely as possible. In particular we want to reproduce the large correlations in this matrix, between Employees and Sales, and between Profits and Cash Flow. Each of these will require a separate factor (column of the  $Q$  matrix), so a solution of at least two factors is probably needed, and we will try a two-factor solution<sup>1</sup>. The `factanal` function does maximum likelihood factor analysis.

<sup>1</sup>The four-factor solution cannot be fit as there are too many parameters relative to the number of variables. The three-factor model can be fit, but but there are as many parameters as there are unique elements in  $R$ .

```
> (f2 <- factanal(uscomp, factor=2, rotation="none", subset=-c(38,40)))
```

Call:

```
factanal(x = uscomp, factors = 2, subset = -c(38, 40), rotation = "none")
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
0.741	0.005	0.413	0.135	0.009	0.362

Loadings:

	Factor1	Factor2
Assets	0.472	-0.189
Sales	0.948	-0.310
MarketValue	0.720	0.263
Profits	0.357	0.859
CashFlow	0.639	0.764
Employees	0.774	-0.199

	Factor1	Factor2
SS loadings	2.775	1.561
Proportion Var	0.462	0.260
Cumulative Var	0.462	0.723

Test of the hypothesis that 2 factors are sufficient.

The chi square statistic is 49.74 on 4 degrees of freedom.

The p-value is 4.09e-10

In the above output:

1. The first argument to `factanal` is in this case the name of a data frame, and by default all columns are used to define  $X$ . You can also specify the columns using a one-sided formula, like `~ Assets + Sales + MarketValue + Profits + CashFlow + Employees`, and using a `data=uscomp` argument. By default the program will convert the sample covariance matrix  $S$  to a correlation matrix before computing. If you want to override this behavior, you can choose the matrix yourself using the `covmat` argument.
2. The *uniquenesses* are the estimates of the diagonal elements of  $\Psi$ . In the textbook, these are called *specific variances*. The larger the specific variance, the less a particular variable is determined by the latent factors. If the uniquenesses are close to 1, then that particular variable is not well “explained” by the common factors. In this example, `assets` is poorly represented by the two common factors, while `CashFlow` is very well represented.
3. The *loadings* are the (an?) estimate of  $Q$ , in this case computed as if  $k = 2$  factors were sufficient. The *communality*, which is one minus the specific variance, is also the row sum of squares  $\sum_j q_{ij}^2$ , and so gives the same information as the specific

variance. If any entries in  $\hat{Q}$  are shown as blank, they are really just *small*: the default is to display a blank if  $|q_{jk}| < .1$ .

The following matrix should approximate  $R$ , if a two-factor solution is adequate:

```
> Q <- loadings(f2)
> Psi <- diag(f2$uniquenesses)
> (sighat <-(Q %*% t(Q) + Psi))
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
Assets	1.000034	0.50626	0.2901	0.006242	0.1573	0.4030
Sales	0.506264	1.00001	0.6010	0.072644	0.3694	0.7954
MarketValue	0.290107	0.60100	1.0000	0.483028	0.6608	0.5046
Profits	0.006242	0.07264	0.4830	1.000003	0.8841	0.1055
CashFlow	0.157339	0.36941	0.6608	0.884076	1.0000	0.3426
Employees	0.402985	0.79535	0.5046	0.105514	0.3426	1.0000

We view percent errors in the estimates of the correlations:

```
> round(100*(sighat - R) / R, 1)
```

	Assets	Sales	MarketValue	Profits	CashFlow	Employees
Assets	0.0	-0.2	-30.0	-116.1	-1.1	93.2
Sales	-0.2	0.0	0.4	0.8	0.0	-0.1
MarketValue	-30.0	0.4	0.0	2.8	-0.1	-22.1
Profits	-116.1	0.8	2.8	0.0	0.0	-40.9
CashFlow	-1.1	0.0	-0.1	0.0	0.0	1.9
Employees	93.2	-0.1	-22.1	-40.9	1.9	0.0

A few of the correlations are not well-approximated, for example between Profits and Assets, and between Employees and Assets, suggesting that the two-factor solution may not be adequate.

4. At the foot of the loadings, the *SS loadings* are the column sum of squares  $\sum_i q_{ij}^2$ .
5. Finally a test is given, with null hypothesis that  $\Sigma$  is of the form (2) versus the alternative that  $\Sigma$  is arbitrary. It can be viewed as a test of dimensionality. The small  $p$ -value suggests that the two-factor model is not adequate. If we try the three-factor model,

```
> (f3 <-factanal(uscomp, factor=3, rotation="none", subset=-c(38,40),
+ scores="regression"))
```

Call:

```
factanal(x = uscomp, factors = 3, subset = -c(38, 40), scores = "regression", rotation = "none", subset = -c(38, 40), scores = "regression")
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
0.567	0.071	0.344	0.106	0.005	0.005

Loadings:

	Factor1	Factor2	Factor3
Assets	0.236		0.611
Sales	0.722	-0.390	0.506
MarketValue	0.803		0.108
Profits	0.647	0.625	-0.292
CashFlow	0.816	0.574	
Employees	0.816	-0.573	

	Factor1	Factor2	Factor3
SS loadings	2.970	1.205	0.728
Proportion Var	0.495	0.201	0.121
Cumulative Var	0.495	0.696	0.817

The degrees of freedom for the model is 0 and the fit was 0.246

We get an exact fit because (2) has as many free parameters as does a general  $\Sigma$ . The two-factor solution is not the first two columns of the three-factor solution. The uniqueness for *Assets* is smaller, but still relatively large.

**Non-uniqueness of estimates and rotations** Interpretation of  $Q$  should allow us to give names to the latent factors. However, if  $G$  is any orthogonal  $k \times k$  matrix, then we can write

$$\begin{aligned}
 x &= QGG'f + u + \mu \\
 &= Q^*f^* + u + \mu \\
 \Sigma &= Q^*Q^{*'} + \Psi \\
 &= (QG)(QG)' + \Psi \\
 &= QQ' + \Psi
 \end{aligned}$$

As long as the distribution of  $f$  is invariant under rotation (for multiplying by an orthogonal matrix is just a rotation of coordinate systems), if  $\hat{Q}$  is an estimate of  $Q$  then so is  $\hat{Q}G$  for any orthogonal  $G$ . Choosing a useful  $G$  is an opportunity to improve interpretable, and a liability, a chance to force the data to agree with a theory. Among the orthogonal  $G$ , a common choice, and the only one we will discuss, is the *varimax rotation*. Now the transformed loadings are given by  $QG$ , and the matrix of their sum of squares and cross-products is  $A = (QG)'(QG)$ , which is a diagonal matrix because  $Q'Q$  is diagonal. The varimax method selects  $G$  to make  $a_{11}$ , the sum of squares of the loadings in the first column, as large as possible,  $a_{22}$  as large as possible subject to the second column of  $QG$  is orthogonal to the first and so on.

```
> (f4 <-factanal(uscomp, factor=3, rotation="varimax", subset=-c(38,40),
+               scores="regression"))
```

Call:

```
factanal(x = uscomp, factors = 3, subset = -c(38, 40), scores = "regression",      rota
```

Uniquenesses:

Assets	Sales	MarketValue	Profits	CashFlow	Employees
0.567	0.071	0.344	0.106	0.005	0.005

Loadings:

	Factor1	Factor2	Factor3
Assets		0.125	0.646
Sales	0.115	0.711	0.641
MarketValue	0.523	0.552	0.278
Profits	0.933		-0.133
CashFlow	0.960	0.188	0.195
Employees	0.132	0.980	0.132

	Factor1	Factor2	Factor3
SS loadings	2.097	1.826	0.979
Proportion Var	0.350	0.304	0.163
Cumulative Var	0.350	0.654	0.817

The degrees of freedom for the model is 0 and the fit was 0.246

For the rotated factors, Factor 1 is a measure of the size of the company, the number of employees, sales and market value. Factor 2 measures money, but not sales, and the remaining factor depends on assets and sales, which might correspond to a theoretical market value of the company<sup>2</sup>.

For each unit in the data there is a vector  $f$  of **factor scores**. Since  $f$  is a random variable, we would speak of predicting  $f$  rather than estimating it. Now

$$\text{Var} \begin{pmatrix} X - \mu \\ f \end{pmatrix} = \begin{pmatrix} \Sigma = QQ' + \Psi & Q \\ Q' & I_k \end{pmatrix}$$

and so the regression prediction, which is justified by multivariate normality of  $x$  and  $f$ , of the factor score is

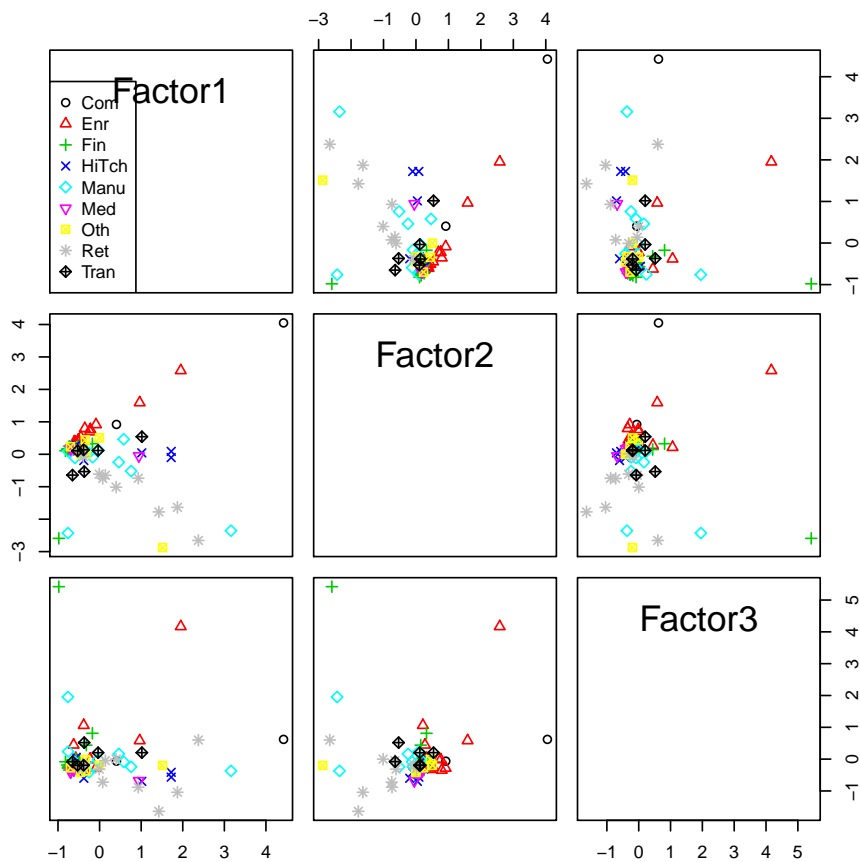
$$E(f|X = x) = Q'S^{-1}(x - \bar{x})$$

where  $S$  is the sample covariance (usually, correlation) matrix. We get the estimated factor scores by the argument `scores="regression"` on the call to `factanal`. Here is a scatterplot:

```
> library(car)
> scatterplotMatrix(f3$scores, group=snames[sector[-c(38,40)]], diagonal="none",
+   reg.line=FALSE, smooth=FALSE)
```

---

<sup>2</sup>In a previous version of this handout, I used `varimax(loadings(f3))` to get the varimax rotation from `f3` rather than by refitting. This works fine except that the factors are not necessarily ordered from most important to least important. Also confusing is the the SS loadings for the varimax solution are smaller than for the "none" solution.



The result is not completely straightforward here. The sector information is helpful. For example the second factor is relatively large for the medical sector.

## Officer ratings

This example consists of fourteen ratings of 103 police officers by their superiors. The data come from the Getting Started page of the SAS help files for `SAS proc factor`.

```
> loc<-"http://www.stat.umn.edu/~sandy/courses/8053/Data/officerratings.csv"
> data <- read.csv(url(loc),header=TRUE)
```

The column names in this data frame are very long, and to improve readability of the output, we will rename them with short names.

```
> (names <- data.frame(vname=paste("Q", 1:14, sep=""),
+ description=names(data)))
```

	vname	description
1	Q1	Communication.Skills
2	Q2	Problem.Solving
3	Q3	Learning.Ability
4	Q4	Judgment.Under.Pressure

```

5     Q5           Observational.Skills
6     Q6 Willingness.to.Confront.Problems
7     Q7           Interest.in.People
8     Q8           Interpersonal.Sensitivity
9     Q9           Desire.for.Self.Improvement
10    Q10          Appearance
11    Q11          Dependability
12    Q12          Physical.Ability
13    Q13          Integrity
14    Q14          Overall.Rating

```

```
> colnames(data) <- names$vname
```

Let's look first at the correlation matrix:

```
> print(R <- cor(data), digits=2)
```

```

      Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9  Q10  Q11  Q12  Q13  Q14
Q1  1.00 0.63 0.55 0.55 0.54 0.53 0.44 0.50 0.56 0.49 0.55 0.22 0.51 0.68
Q2  0.63 1.00 0.57 0.62 0.43 0.50 0.40 0.44 0.41 0.39 0.45 0.32 0.38 0.58
Q3  0.55 0.57 1.00 0.49 0.62 0.52 0.27 0.19 0.57 0.40 0.51 0.23 0.31 0.59
Q4  0.55 0.62 0.49 1.00 0.37 0.40 0.62 0.61 0.48 0.23 0.55 0.35 0.59 0.66
Q5  0.54 0.43 0.62 0.37 1.00 0.73 0.26 0.17 0.60 0.42 0.56 0.43 0.39 0.58
Q6  0.53 0.50 0.52 0.40 0.73 1.00 0.22 0.13 0.53 0.48 0.49 0.49 0.33 0.59
Q7  0.44 0.40 0.27 0.62 0.26 0.22 1.00 0.81 0.49 0.27 0.61 0.38 0.75 0.61
Q8  0.50 0.44 0.19 0.61 0.17 0.13 0.81 1.00 0.37 0.26 0.54 0.22 0.69 0.58
Q9  0.56 0.41 0.57 0.48 0.60 0.53 0.49 0.37 1.00 0.45 0.60 0.38 0.57 0.67
Q10 0.49 0.39 0.40 0.23 0.42 0.48 0.27 0.26 0.45 1.00 0.51 0.38 0.41 0.57
Q11 0.55 0.45 0.51 0.55 0.56 0.49 0.61 0.54 0.60 0.51 1.00 0.45 0.65 0.77
Q12 0.22 0.32 0.23 0.35 0.43 0.49 0.38 0.22 0.38 0.38 0.45 1.00 0.38 0.44
Q13 0.51 0.38 0.31 0.59 0.39 0.33 0.75 0.69 0.57 0.41 0.65 0.38 1.00 0.67
Q14 0.68 0.58 0.59 0.66 0.58 0.59 0.61 0.58 0.67 0.57 0.77 0.44 0.67 1.00

```

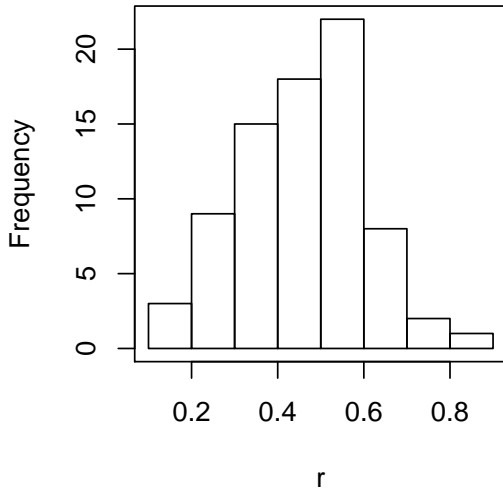
```

> par(mfrow=c(1, 2))
> r <- R[-14, -14][lower.tri(R[-14, -14])]
> hist(r, main="Sample Correlations", xlab="r")
> box()
> require(car)
> z <- 0.5 * log((1 + r)/(1 - r))
> qqPlot(z, main="Fisher's z-transform")

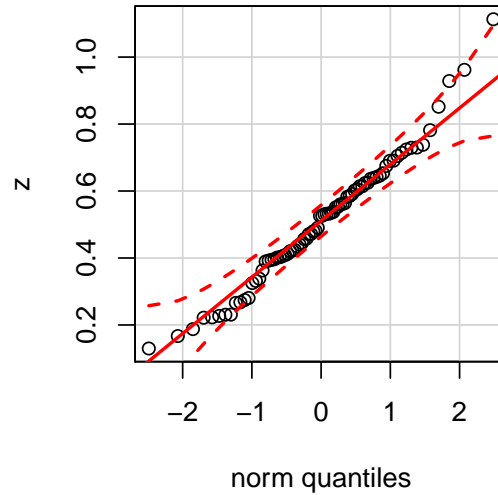
```

If all the correlations really equal, then the Fisher's  $z$ -transform should be approximately  $N(\rho, 1/(n-3))$ . The sd of the Fisher  $z$ -transforms is 0.189, as compared to  $\sqrt{1/(n-3)} = 0.1$ .

**Sample Correlations**



**Fisher's z-transform**



All the elements are positive and most are in the range 0.4 to 0.6. The structure here is unclear.

```
> (f3 <- factanal(~.-Q14, data=data, factors=3, rotation="varimax"))
```

Call:

```
factanal(x = ~. - Q14, factors = 3, data = data, rotation = "varimax")
```

Uniquenesses:

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
0.383	0.253	0.419	0.368	0.261	0.323	0.181	0.168	0.408	0.651
Q11	Q12	Q13							
0.335	0.689	0.267							

Loadings:

	Factor1	Factor2	Factor3
Q1	0.455	0.356	0.533
Q2	0.303	0.256	0.768
Q3	0.588		0.478
Q4	0.269	0.554	0.503
Q5	0.825		0.226
Q6	0.757		0.321
Q7	0.186	0.873	0.148
Q8		0.867	0.282
Q9	0.644	0.377	0.188
Q10	0.515	0.211	0.197
Q11	0.573	0.548	0.188
Q12	0.488	0.265	
Q13	0.372	0.765	0.101

	Factor1	Factor2	Factor3
SS loadings	3.389	3.173	1.733
Proportion Var	0.261	0.244	0.133
Cumulative Var	0.261	0.505	0.638

Test of the hypothesis that 3 factors are sufficient.  
The chi square statistic is 63.39 on 42 degrees of freedom.  
The p-value is 0.0181

```
> (f4 <- factanal(~.-Q14, data=data, factors=4, rotation="varimax"))
```

Call:

```
factanal(x = ~. - Q14, factors = 4, data = data, rotation = "varimax")
```

Uniquenesses:

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
0.356	0.144	0.390	0.380	0.259	0.325	0.179	0.166	0.386	0.647
Q11	Q12	Q13							
0.339	0.039	0.275							

Loadings:

	Factor1	Factor2	Factor3	Factor4
Q1	0.556	0.401	0.413	
Q2	0.360	0.263	0.803	0.114
Q3	0.683	0.128	0.356	
Q4	0.302	0.570	0.436	0.122
Q5	0.828	0.114	0.106	0.176
Q6	0.735		0.239	0.276
Q7	0.144	0.871	0.122	0.162
Q8		0.879	0.247	
Q9	0.659	0.403		0.108
Q10	0.493	0.217	0.149	0.202
Q11	0.549	0.560	0.110	0.183
Q12	0.284	0.206		0.913
Q13	0.336	0.769		0.137

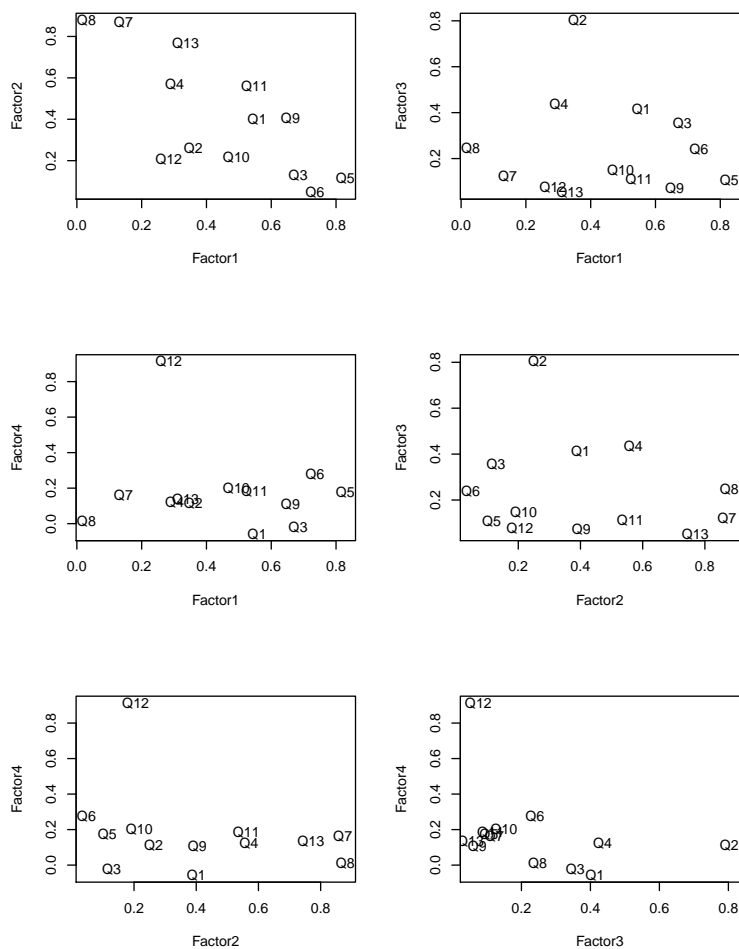
	Factor1	Factor2	Factor3	Factor4
SS loadings	3.415	3.274	1.325	1.103
Proportion Var	0.263	0.252	0.102	0.085
Cumulative Var	0.263	0.515	0.616	0.701

Test of the hypothesis that 4 factors are sufficient.  
The chi square statistic is 40.08 on 32 degrees of freedom.  
The p-value is 0.154

The three-factor solution is inadequate, while the four-factor solution provides a reasonable approximation to the correlation matrix, explaining about 70% of the variability.

Look first at the uniquenesses. Q10, Appearance, has a very large uniqueness, suggesting that it is largely not determined by the common factors; for the other questions the uniquenesses are somewhat smaller. Let's look at the loadings in graphs:

```
> par(mfrow=c(3,2))
> loads <- loadings(f4)
> for (j in 1:6) {
+   h <- c(1,1,1,2,2,3)[j]
+   v <- c(2,3,4,3,4,4)[j]
+   plot(loadings[,c(h,v)],type="n")
+   text(loadings[,c(h,v)],rownames(loadings)) }
```



```
> f4 <- update(f4,scores="regression")
> pairs(f4$scores)
```

